

Fully Convolutional Architectures for Multi-Class Segmentation in Chest Radiographs

Alexey A. Novikov, David Major, Dimitrios Lenis, Jiří Hladůvka, Maria Wimmer, and Katja Bühler

Abstract—The recent success of Deep Convolutional Neural Networks on image classification and recognition tasks has led to new applications in very diversifying contexts. One of these is medical imaging where scarcity and imbalance of training data has hindered rapid development of neural network related applications.

This paper investigates and proposes neural network architectures within the context of automated segmentation of anatomical organs in chest radiographs, namely for lung, clavicles and heart. By relating prior class data distributions to the objective function sparsely represented structures are methodologically emphasized. Scarce training sets and data augmentation are encountered with aggressive data regularization. The problem of highly imbalanced target object appearance in the input data is solved by modifying the objective function.

The models are trained and tested on the publicly available JSRT database consisting of 247 X-Ray images the ground-truth masks for which available in the SCR database. The networks have been trained in a multi-class setup with three target classes.

Our best performing model trained with the negative Dice loss function was able to reach mean Jaccard overlap scores of 94.1% for lungs, 86.6% for heart and 88.7% for clavicles in the multi-label setup, therefore, outperforming the best state-of-the-art methods for heart and clavicle and human observer on lung and heart segmentation tasks.

Index Terms—Lung segmentation, clavicle segmentation, heart segmentation, fully convolutional network, regularization, imbalanced data, chest radiographs, multi-class segmentation, JSRT dataset

I. INTRODUCTION

DESPITE a plethora of modalities and their combinations in current state of the art medical imaging, radiography holds an esteemed position, forming together with ultrasonography the two main pillars of diagnostic imaging, helping in solving between 70-80% of diagnostic questions [1]. Considering posterior anterior chest radiographs (CXR) as an example, their importance is apparent: in an NHS technical report [2] they are listed as the leading tool in diagnosis and treatment; variations of the size, positions and areas of heart, lung fields, hila structures, clavicles etc. may give indications on the presence of TBC, cancer and other diseases or assist in their early diagnosis.

Hence semantic segmentation of radiographs has been an active field of study. Individual anatomical intricacies like high interpersonal variations in shape and size of central organs like lung fields, clavicles and heart, related to age, size and gender, ambiguous organ boundaries due to organ overlaps and

artifacts caused by movements and image modality intrinsics, are just a few of the reasons why accurate organ segmentation on CXR images still remains an inherently challenging task.

Unsurprisingly solutions therefore vary in their favored toolsets including *rule-*, *shape-* and *graph-* based methods, *pixel classifications* and *statistical approaches*. Based on their recent, rapid success in computer vision challenges [3] *neural network* (NN) based approaches are successfully used. These solutions, typically avoid using task-specific, manually engineered features. Such features tend to be more capable of describing different aspects of an object. Usually they derive their generalizability out of vast datasets (e.g. ImageNet database [4] holds more than one million labeled training images) that are not available in the medical domain, thereby indicating the need for more delicately assembled network structures.

A. Related Work

Due to their above-stated significance, semantic CXR segmentation has been studied extensively in the literature. As approaches vary vastly, only solutions which had been tested against a common dataset are listed below. These training and test sets consist of CXRs made available through the Japanese Society of Radiological Technology (JSRT) [5]. The SCR database with manual segmentations for lung fields, heart and clavicles was introduced in the study by Ginneken et al. [6]. In the algorithms compared in the study the authors performed their evaluations on downsampled images of the smaller resolution, namely 256×256 , which reduces computational complexity, but due to downsampling introduces an additional, in specific use cases maybe even problematic border smoothing. As a common overlap measure the different groups used the *Jaccard Index* [7].

Classical Approaches: Following and augmenting van Ginneken et al. [6], the space of algorithmic approaches may be roughly partitioned into *rule-*, *shape-* and *graph-* based methods, *pixel classification* and *statistical approaches*. Each methodological framework has their own set of advantages, e.g. by limiting to a predefined rule-set or deformable shape, rule and shape based methods will yield anatomical sound solutions. Graph-based methods build upon the anatomy inherent topology and therefore will also adhere to that principle, but simultaneously allow for a higher class of variations, with the tradeoff of higher computational complexity. Pixel classification and statistical approaches treat the problem as a local classification / optimization task and therefore allow for higher variations, maintaining a traceable computation at

A. A. Novikov, D. Major, D. Lenis, J. Hladůvka, M. Wimmer and K. Bühler are with the VRVis Center for Virtual Reality and Visualization, 1220 Vienna, Austria, e-mail: (novikov@vrvis.at, major@vrvis.at, lenis@vrvis.at, hladuvka@vrvis.at, mwimmer@vrvis.at, buehler@vrvis.at).

the cost of sometimes unrealistic outcome. Typically as the following show, best results can be achieved by hybrid approaches, i.e. the combination of efficient initial segmentations with successive detailed adoptions to plausible outputs.

Adhering to their importance *lung fields* have received significant attention. Candemir et al. [8] developed a registration-based algorithm. The same group proposed a graph-cut based algorithm [9]; focusing on (smoothing-) parameter adaptation, the algorithm achieved a mean score of 0.976 promising qualitatively robust results and time-efficient calculation. A recent hybrid approach stems from the work by Shao et al. [10], combining active shape and appearance models; the group yields a comparable overlap score of 0.946. Ibragimov et al. [11] demonstrate how the combination of landmark-based segmentation and random shape based random forest classifier can achieve an overlap score of 0.953, while still maintaining computational efficiency. Based on an active shape model approach, specifically addressing the initialization dependency of these models [12], achieves an improvement and an overlay score of 0.954. van Ginneken et al. [6] survey older approaches, back to 2006, that score on the same dataset comparably between 0.713 and 0.969; the survey also evaluated a human observer with a score of 0.9468 which did not vary statistically significantly from the survey leading pixel classification method. Overall lung field segmentation in CXR remains an active topic, with algorithmic setups rivaling the human observer.

In comparison, segmentation of *clavicles* has shown to be more challenging. High variations of their positioning and general shape, the impact of bone density on the radiograph and their overlap with rib and lung structures yield in high person-specific intricacies, thereby big deviations from a theoretical average clavicle and hence a steep impact on overlay scores. van Ginneken et al. [6] include clavicle segmentation in the survey where they were able to reach scores between 0.505 and 0.734. The human observed reached 0.896 which demonstrates that comparing to segmentation of lung and heart fields this task is challenging even for humans. Hogeweg et al. [13] developed a combination of pixel classification, active shape model and dynamic programming that led them to an overlap of 0.850; however the overlap was only measured within the lung fields. Predominantly shape/contour- based models vary on the choice of the underlying feature space. Exemplified on the approach presented by Boussaid et al. [14], the problem is addressed as a deformable contour model, that uses SIFT features to describe the embedding object appearance. Using this, they achieved an overlap score of 0.904.

Starting from their respective lung field segmentation, most approaches are geared towards a generalizable solution, trying to adapt their algorithm to also span the *hearts* requirements. Similarly, van Ginneken et al. [6] and Boussaid et al. [14] also report on their segmentation results, yielding in between 0.77 and 0.86 and mean 0.91 overlap scores respectively. In a recent study, Candemir et al. [15] specifically adapted their registration based method [8] to fit heart-position extraction, and achieved an overlap accuracy between 0.697 and 0.087 (compared to 0.954 for their lung field segmentation). Generally, as the heart boundaries are overlapped and occluded by

the surrounding lung fields, hence are not clearly visible, exact segmentation remains challenging.

Neural networks: While conceptually more than 50 years old Neural Networks (NN), the abstracted basis of *deep learning*, are living through a revival [3]. A deeper understanding of training and numerical behavior and the steep increase of tractable calculation schemes through the leveraging of graphical processing units (GPUs) has allowed this class of approach to become the de facto standard, or at least serious contender in several machine learning branches [3], [16]. For brevity, the following focuses on convolutional neural networks (CNNs), successfully used subclass of NN in computer vision tasks [17].

A prototypical setup of such CNNs consists of a combination of convolution filters, interspersed with data reduction and pooling layers [18]. The driving idea is to mimic human visual cognition, in that sense that the complete picture is derived out of low-level features, e.g. edges and circles, which in return yield more distinctive features and finally the desired target through recombination in each successive layer.

Regarding the segmentation of medical images several such setups have been studied; e.g. Greenspan et al. [18] made a summary of the recent state-of-the-art works in the area. As stated above *semantic segmentation* typically builds upon a vast set of training data, e.g. ImageNet [4] and Pascal VOC-2012 [19]. Such large datasets are not typical for the medical domain, rendering most current approaches unfeasible, hence calling for a finely tailored strategy. First attempts date back more than 15 years ago; Tsujii et al. [20] use a NN for lung field segmentation yielding in accuracy around 86%. Aece et al. [21] use a CNN as a binary classifier and thereby partition chest radiographs into the two {bone, non-bone} sets in a fully-automated fashion. NNs do not need to be considered as a standalone solution as was demonstrated by Ngo et al. [22]. The group combined regularized level sets with a deep learning approach and yielded on JSRT overlap scores between 0.948 and 0.985. While CXR segmentation has not been covered extensively yet, different modalities like ultrasound, CT and MRT have been explored [23], [24], [25], [26].

Long et al. [27] address the need for local features that coincide with global structures, and define the *Fully Convolutional Net*. This type of network allows for arbitrary sized input and output. In combination with layer fusion, i.e. shortcuts between selected layers, this setup achieves a nonlinear, local-to-global feature representation, and allows for a pixelwise classification. By adapting this network-class with successive upsampling layers, i.e. enlarging the field of view of the convolution, Ronneberger et al. [28] guide the resolution of feature extraction, and thereby control the local-to-global relations of features. The authors used excessive data augmentation by applying elastic deformations in order to cope with the lack of training data for the cell segmentation task. Elastic deformations however are not reasonable in case of CXR images because that would make rigid organs such as lungs, heart and clavicles look anatomically incorrect and moreover could confuse training by making the network learn features corresponding to such unrealistic structures.

We explicitly address the issues of *limited options regarding*

non-rigid transformations, unbalanced organ representation and ambiguous organ boundaries. We specifically adapt the U-Net model [28] for CXR images by applying other regularization methods and building new architectures performing successfully without additional data augmentation. We also propose and compare two different training loss functions to deal with the problem of the multi-class segmentation in the case of imbalanced data representation which is the case for clavicles in CXR images as they are under-represented in the sense of pixel area comparing to heart and lung fields. Our solution allows us to simultaneously yield overlap scores, comparable and in many cases surpassing state-of-the-art techniques and human performance on all considered tasks including the classically challenging cases of heart and clavicle segmentation.

B. Contributions

The major contributions of our work include the following:

- 1) We propose a multi-class end-to-end approach for segmentation of anatomical organs in chest radiographs.
- 2) We propose a solution for training the fully convolutional models in case of imbalanced data representation.
- 3) We propose architectures which outperform state-of-the-art methods by a large margin, especially on the clavicle segmentation task on publicly available JSRT dataset.

II. METHODOLOGY

In this section, we begin with a formal description of the multi-class approach. We then shortly describe the base setup our methods are built on and, finally, we outline our architecture design strategies and propose a number of models applicable to solving the problems specific for CXR images.

A. Multi-Class Approach

The input data consists of a set of 2D images $\mathcal{J} = \{I \mid I \in \mathbf{R}^{m_1 \times m_2}\}$ and the corresponding multi-channel binary ground-truth masks $(L_{i,I})_{1 \leq i \leq n}$ where $L_i \in \{0,1\}^{m_1 \times m_2}$, n is the number of classes we aim to address, and m_1, m_2 are the image dimensions.

We first split \mathcal{J} into sets $\mathbf{I}_{\text{TRAIN}}$ of size $K = |\mathbf{I}_{\text{TRAIN}}|$ and $\mathbf{I}_{\text{TEST}} = \mathcal{J} \setminus \mathbf{I}_{\text{TRAIN}}$. As described above, for each $I \in \mathcal{J}$ a series of binary ground-truth masks $(L_{i,I})_{1 \leq i \leq n}$ is used. For a later reference let \mathcal{L} be the set of all ground truth classes, hence $1 \leq n \leq |\mathcal{L}|$.

The networks are trained in the following manner: the network is consecutively passed with minibatches $\mathcal{K} \in \mathcal{N}$ where \mathcal{N} is a partition of the set $\mathbf{I}_{\text{TRAIN}}$. Minibatches \mathcal{K} are non-empty subsets of $\mathbf{I}_{\text{TRAIN}}$ derived in a way that every image $I \in \mathbf{I}_{\text{TRAIN}}$ is included in one and only one of the minibatches \mathcal{K} . Additionally, we introduce $c_{\mathcal{K}}$ to define the total pixel count over all $I \in \mathcal{K}$.

For each $I \in \mathcal{K}$ the multi-class output of the network is calculated, i.e. understanding the network as a function

$$\mathcal{F} : \mathcal{J} \rightarrow (\{0,1\}^{m_1 \times m_2})_{1 \leq i \leq n} \quad (1)$$

Therefore, for each pixel of I its semantic class $l \in \mathcal{L}$ can be derived in a single step up to some probability.

In order to estimate and maximize this probability we can define an energy function

$$\Lambda_{(L_{i,I})} : \{0,1\}^{m_1 \times m_2} \times (L_{i,I}) \rightarrow \mathbf{R} \quad (2)$$

that estimates the deviation (error) of the network outcome from the desired ground-truth. The error is back-propagated then to update the network parameters. The whole procedure continues until the defined given stopping criteria are fulfilled.

At testing time an unseen image $I \in \mathbf{I}_{\text{TEST}}$ is passed through the network and the multi-label output $\mathcal{F}(I)$ is produced. As defined above, the network output consists of series of multi-channel segmentation masks. The channels in case of chest radiographs correspond to different body organs.

The model is built, initialized and further trained. After the training is finished the learnt model weights and regularization layers are fixed and the model is validated on a set of test images. Main steps of the method are introduced in the following sections in detail.

B. Base Setup

The U-Net like architecture which was originally proposed by Ronnenberger et al. [28] consists of contraction and expansion parts. In the contraction part high-level abstract features are extracted by consecutive application of pairs of convolutional and pooling layers. In the expansion part the low-level abstract features are merged with the features from the contractive part respectively. The output of the network is a multi-channel segmentation mask where each channel has the same size as the input image.

Excellent performance of the Original U-Net architecture has been demonstrated for segmentation of neuronal structures in electron microscopic stacks [28]. For other subject-specific tasks it however requires additional modifications due to a different data representation. In particular, when data is highly imbalanced or in cases when data augmentation is not reasonable. The problem on imbalanced data in medical images occurs due to different sizes of anatomical organs of interest. For example, in JSRT dataset ground-truth masks 60% of pixels belong to background, 29% to lung, 2% to clavicles and 9% to heart respectively, hence emphasizing lung and heart fields over clavicles.

C. Improvements of U-Net Model for Chest Radiographs

On top of the original architecture we analyze and evaluate the network by introducing a number of modifications in *architecture* and *training*. We consider a number of possible improvements of the network model in detail and based on the evaluation results propose a number of models tailored to efficiently train and perform multiclass segmentation on medical CXR images. To avoid the data augmentation which was applied in the method by Ronneberger et al. [28] we propose to modify the model by using a more aggressive regularization. On top of this we propose a number of architectures to further improve the segmentation result. In addition to a different model regularization and architectural modifications we propose a different training loss function strategy to cope with the problem of highly imbalanced data

representation. An overview of the proposed architecture is depicted in Fig. 1.

Architectural Modifications:

Acquiring more training data would be of benefit for any learning algorithm. However, in medical imaging getting additional data is not always feasible.

Ronneberger et al. [28] used elastic deformations for data augmentation in order to regularize the model. However elastic deformations are not reasonable in case of chest radiographs because they would make rigid organs such as lungs, heart and clavicles look anatomically incorrect and could then confuse training by making the network learn features corresponding to unrealistic structures.

The number of feature maps and layers in the original version of U-Net is large which results in tens of millions of parameters in the system which slows down training and does not necessarily decrease generalization error. Without any regularization training of such large networks can *overfit on the data*. Especially when there is not much training data available. Overfitting is especially a problem for smaller or thinner prolonged anatomical organs such as clavicles due to their more varying shape representations in CXR images. In the case when the network architecture is deep and availability of training data is limited, another possibility to decrease the generalization test error of the algorithm is more aggressive regularization.

a) All-Dropout - improving accuracy with aggressive regularization: Dropout layer [29] has become a common practice in modern deep network architectures. Moreover, according to Bouthillier et al. [30] it can also play a role of data augmentation at the same time. We therefore propose an architecture with a dropout layer after every convolutional layer in the network. We use the Gaussian dropout which is equivalent to adding a Gaussian distributed random variable with zero mean and standard deviation equal to the activation of the neural unit. According to Srivastava et al. [29] it works even better than the classic one which uses the Bernoulli distribution. Besides, adding such noise is a more natural choice for chest radiographs due to Gaussian noise occurring during their acquisition [31]. In the following we address this architecture as *All-Dropout*.

b) InvertedNet - improving accuracy with fewer parameters: One way of dealing with model overfitting is to reduce the number of parameters. We propose a modification of the *All-Dropout* architecture by a) performing the delayed subsampling of the first pooling layer with (1,1) pooling and b) changing the numbers of feature maps in the network. In this architecture we propose to start with the large number of feature maps and divide it by the factor of two after every pooling layer and increase by the factor of two after every upsampling layer respectively. In this case the networks learn many different variations of structures at the early layers and less high level features at the later layers. This seems more reasonable in case of more rigid anatomical organs such as clavicles because their shapes in the end do not vary too

much and therefore there is no need to learn too many high abstract features. We will call this architecture *InvertedNet* due to the way the numbers of feature maps are changed with respect to the Original U-Net architecture.

c) All-Convolutional - improving accuracy by learning pooling: Springenberg et al. [32] showed that having pooling layers replaced by convolutional layers with a higher stride or removing pooling layers completely can improve final results. This modification introduces new parameters in the network but can be considered as a learning of pooling for each part of the network rather than just fixing pooling parameters to constant values. Such pooling learning can be useful for learning better features for smaller and thinner elongated objects. Further motivated the work by Springenberg et al. [32] we adapt the Original U-Net accordingly to the model All-CNN-C: each pooling layer is replaced by a convolutional layer with filter size equal to the pooling size of the replaced pooling layer. We modify the *All-Dropout* architecture correspondingly and further call this architecture *All-Convolutional*.

Training Strategies:

As already mentioned above large differences in sizes between anatomical organs of interest can introduce a problem of imbalanced data representation. In such cases classes are represented in highly different amounts pixel-wise and therefore losses for sparsely represented classes can go unnoticed sometimes. Hence, classic formulations of loss such as cross-entropy or negative dice functions would underestimate the classes represented in very small amounts. We address the imbalance in pixel representation by introducing a weighted distance function.

Following the section II-A let \mathcal{L} be the set of all ground-truth classes and \mathcal{N} a partition of our training set. For $\mathcal{K} \in \mathcal{N}$ and $c_{\mathcal{K}}$ its total pixel count we define $r_{\mathcal{K},l}$ as the ratio:

$$r_{\mathcal{K},l} := \frac{c_{l,\mathcal{K}}}{c_{\mathcal{K}}} \quad (3)$$

where $c_{l,\mathcal{K}}$ is the number of pixels belonging to the semantic class $l \in \mathcal{L}$ in the training batch \mathcal{K} .

For a distance function $d : \{0,1\}^{m_1 \times m_2} \times \{0,1\}^{m_1 \times m_2} \rightarrow \mathbb{R}$, and an image $I \in \mathcal{K}$ we minimize our objective function

$$\Lambda_{(L,I)}(I) := \sum_{l \in \mathcal{L}} r_{\mathcal{K},l}^{-1} d(\mathcal{F}(I)_l, L_{l,I}) \quad (4)$$

over the set \mathcal{K} and the complete partition. By this sparsely represented classes, e.g. clavicles, are no longer under-represented in favor to large ground-truth mask, e.g. lung fields.

For d we chose and evaluated the so-called *weighted pixel-wise cross-entropy* and *weighted negative dice* loss functions. Cross-entropy is a typical choice for neural networks and dice seems to be a natural choice in case of segmentation problems. The *weighted dice loss* function in our case takes the sigmoid activation computed at the final output feature map for each channel as the input. The sigmoid activation is defined as:

$$p_k(\mathbf{x}) := \frac{1}{1 + e^{-a_k(\mathbf{x})}} \quad (5)$$

where $\mathbf{a}_k(\mathbf{x})$ indicates activation at feature channel k at the pixel $\mathbf{x} \in I$ and $p_k(\mathbf{x})$ is the approximated probability of the pixel \mathbf{x} not belonging to background. In the case of *weighted negative dice* the output does not have to provide a channel for the background class.

Given an image I , let $\{L_i\}$ be the set of non-background pixels in the corresponding ground-truth multi-channel mask and $P_k(I)$ is the set of pixels where the model is sure that they do not belong to the background:

$$P_k(I) := \{\mathbf{x} : \mathbf{x} \in I \wedge |p_k(\mathbf{x}) - 1| < \epsilon\} \quad (6)$$

where ϵ is a small tolerance value.

The distance function d for the negative Dice coefficient for a training image I can then be defined as:

$$d_{L_k, I}(I) := -2 \frac{|P_k(I) \cap \mathcal{G}_k(I)|}{|P_k(I)| + |\mathcal{G}_k(I)|} \quad (7)$$

where $P_k(I)$ is the predicted segmentation mask and $\mathcal{G}_k(I)$ is the corresponding ground-truth segmentation mask for the image I for the channel k .

The *weighted pixelwise cross-entropy* takes the softmax activation computed at the final output feature map for each channel as the input. We use the softmax $p_k(\mathbf{x})$ as defined by Ronneberger et al. [28] and the distance function d of the cross-entropy for a training image I we define as:

$$d_{L_k, I}(I) := \sum_{\mathbf{x} \in I} \frac{\mathbb{1}_{\mathcal{G}_k(I)} \log p_k(\mathbf{x})}{c_K} \quad (8)$$

D. Proposed Network Architectures

We summarize the proposed architectures:

- *All-Dropout*: Modified version of the U-Net architecture [28] with dropout layers placed after every convolutional layer. Depicted at Fig. 1a.
- *InvertedNet*: Similarly to *All-Dropout* with the delayed subsampling of the first pooling layer and numbers of feature maps in the network inverted with respect to Original U-Net. Depicted at Fig. 1b.
- *All-Convolutional*: Similar to *All-Dropout* with pooling layers replaced by new convolutional layers with filter sizes equal to the pooling size of the corresponding pooling layer. Depicted at Fig. 1c.

We use padded convolutions in all architectures for all convolutional layers. Therefore output channels will have the same size as the input image except the *J-Net* architecture where input dimensions are four times larger than of the output. All proposed architectures contain convolutional and dropout layers. In all architectures all convolutional layers are followed by dropout layers except the third convolutional layers in the *All-Convolutional* architecture where the layer plays a role of the pooling layer it replaces. In all models we used the rectified linear unit functions [33] at all convolutional hidden layers. It is the most common and well performing activation function in modern network architectures [3].

To reduce the number of parameters and speed up training, instead of the last dense layers we used the convolutional layer with the number of feature maps equal to the number of considered classes in case of the weighted dice and with one more for background in case of weighted pixelwise cross-entropy functions. To splash the values to the $[0, 1]$ range at the output of the network we used the sigmoid function as an activation at the output layer.

III. EXPERIMENTS

A. JSRT dataset

We use the JSRT dataset [5] both for training and testing. The dataset consists of 247 posterior anterior (PA) chest radiographs with a resolution of 2048×2048 , 0.175 mm pixel size and 12 bit depth. The reference organ boundaries for JSRT images for left and right lung fields, heart and left and right clavicles were introduced by van Ginneken et al. [6] in 1024×1024 resolution and available in the SCR database.

B. Training Model

Data has been normalized using the mean and standard deviation across the whole training dataset. It has been zero-centered first by subtracting the mean and then normalized additionally by scaling using its standard deviation. It was then split into training and testing sets. Models were trained on images of one of the following resolutions: 128×128 , 256×256 and 512×512 . Original images and masks were downsampled to these resolutions by the local averaging algorithm. To make the paper comparable with state-of-the-art methods, most results in our work correspond to the 256×256 image resolution.

To optimize the model we used the Adaptive Moment Estimation method (ADAM) [34] as it employs an adaptive learning rate approach for each parameter. It stores decaying average of both past squared gradients and past gradients. We varied different initial learning rates in order to find the most stable convergence and 10^{-5} and $5 * 10^{-5}$ seemed to be the most reasonable choices. Training converged slower in the former but more stable than with the latter one. We therefore used the fixed initial rate of 10^{-5} in all our experiments. We used the early stopping criterion with 200 epochs to avoid overfitting.

C. Performance Metrics

To evaluate the architectures and compare with state-of-the-art works, we used the following performance metrics:

Dice Similarity Coefficient:

$$D = \frac{2 \times |G \cap S|}{|G| + |S|} \quad (9)$$

Jaccard Similarity Coefficient:

$$J = \frac{|G \cap S|}{|G| + |S| - |G \cap S|} \quad (10)$$

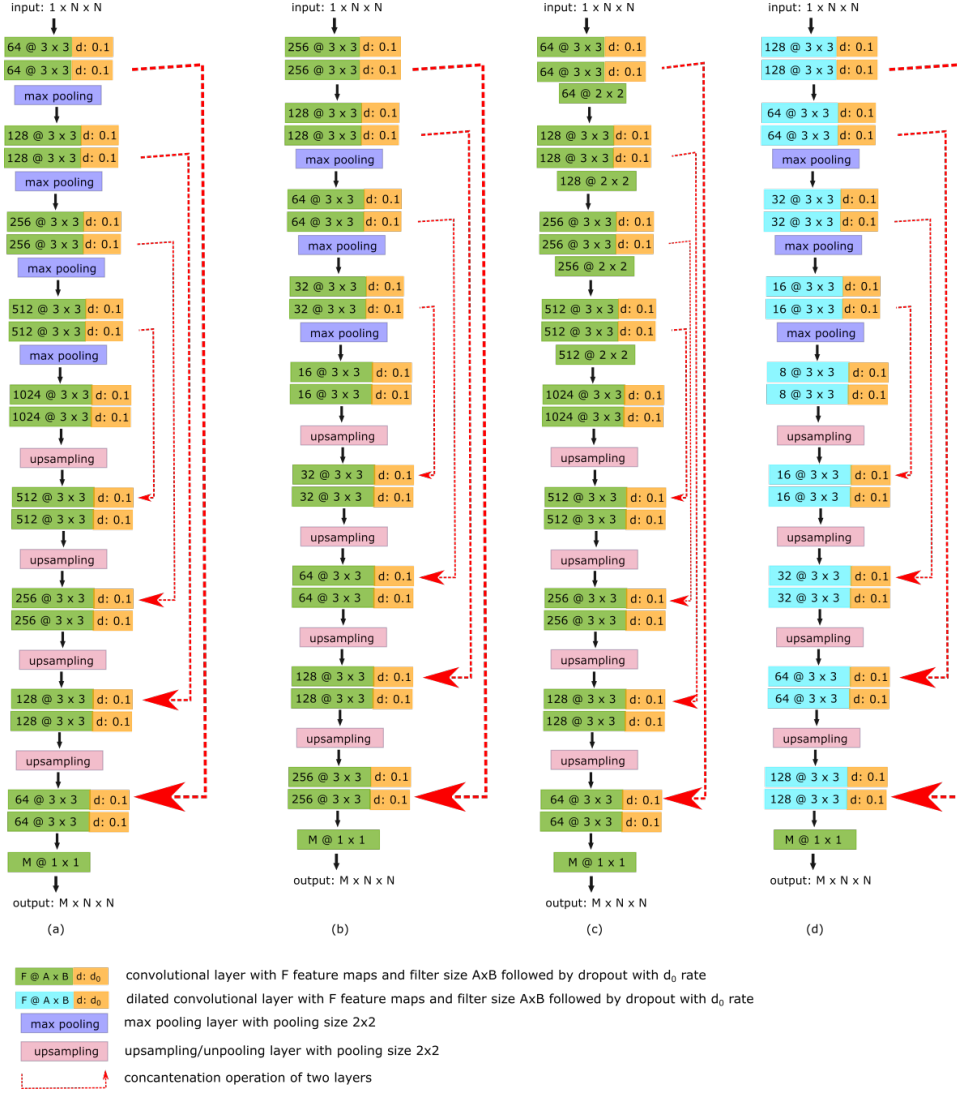


Fig. 1: Proposed Network Architectures a) All-Dropout b) InvertedNet c) All-Convolutional d) InvertedNet with dilated convolutions

Bodypart	Lungs			Clavicles			Heart		
Evaluation Metric	D_{si}	J	S_{sg}	D_{si}	J	S_{sg}	D_{si}	J	S_{sg}
Human Observer [6]		0.946	1.64	-	0.896	0.68	-	0.878	3.78
PC post-processed [6]	-	0.945	1.61	-	0.615	2.90	-	0.824	5.20
Hybrid Voting [6]	-	0.949	1.62	-	0.736	1.88	-	0.860	4.24
ASM Tuned [6]	-	0.927	2.30	-	0.734	2.04	-	0.814	5.96
Original U-Net 128	0.964	0.930	0.98	0.831	0.711	2.58	0.934	0.876	2.19
All-Convolutional 128	0.966	0.934	0.92	0.826	0.703	1.98	0.929	0.867	2.30
All-Dropout 128	0.966	0.934	0.90	0.841	0.725	1.95	0.930	0.870	2.27
InvertedNet 128	0.969	0.941	0.81	0.887	0.796	1.24	0.940	0.887	2.02
Original U-Net 256	0.968	0.939	0.87	0.854	0.740	2.05	0.935	0.877	2.03
All-Convolutional 256	0.970	0.941	0.84	0.864	0.761	1.88	0.937	0.882	1.96
All-Dropout 256	0.971	0.943	0.84	0.875	0.778	1.70	0.939	0.885	1.92
InvertedNet 256	0.972	0.945	0.76	0.896	0.812	1.52	0.928	0.865	2.07
Original U-Net 512	0.972	0.945	0.77	0.873	0.774	2.51	0.926	0.862	1.82
All-Convolutional 512	0.974	0.949	0.71	0.891	0.804	3.00	0.933	0.875	1.72
All-Dropout 512	0.972	0.945	0.75	0.894	0.808	1.54	0.931	0.871	1.73
InvertedNet 512	0.958	0.919	1.08	0.807	0.677	4.88	0.907	0.830	2.11
InvertedNet ¹ 512	0.963	0.929	0.96	0.906	0.829	1.79	0.909	0.834	1.93

TABLE I: Evaluation comparison of proposed models vs. state-of-the-art methods; ⁽¹⁾ corresponds to the architecture depicted at Fig. 1d

where in both coefficients D_{si} and J , G represents the ground-truth data and S stands for the segmentation provided by the evaluated method.

Symmetric Mean Absolute Surface Distance:

$$S_{sg} = \frac{1}{(n_s + n_g)} \times \left(\sum_{i=1}^{n_s} |d_i^{sg}| + \sum_{j=1}^{n_g} |d_j^{gs}| \right) \quad (11)$$

where n_s is the number of pixels in the segmentation provided by the evaluated method, n_g is the number of pixels in the ground-truth data mask, d_i^{sg} is the distance from i -th pixel in the segmentation to the closest pixel in the ground-truth data mask, and d_j^{gs} is the distance from j -th pixel in the ground-truth data mask to the closest pixel in the segmentation provided by the evaluated method.

IV. RESULTS AND DISCUSSION

A. Segmentation Performance

Evaluation results for the proposed architectures for different resolutions are shown in Table I. In addition, results for Original U-Net for three resolutions as well as the best performing methods and human observer results introduced by van Ginneken et al. [6] are added for comparison. Table I is subdivided into five blocks. The first block contains only the human observer result. The second block contains results for the best performing methods summarized by van Ginneken et al. [6]. The third, fourth and fifth blocks contain results of the Original U-Net and the proposed architectures for three different resolutions. Best results for each metric at each block are highlighted in bold font.

Scores for *lung segmentation* did not vary significantly across the methods. All approaches were able to reach results close to human performance. Though none of our architectures actually outperformed human observed and the Hybrid Voting method [6], one of the models reached the same Jaccard score and all of proposed architectures as well as Original U-Net achieved more accurate object contours according to the symmetric surface distance.

Clavicle segmentation was a bit more challenging task for all our architectures. And it is not surprising because clavicles are much smaller than heart and lungs and their shapes change more significantly from one scan to another. None of the proposed methods could outperform human observer though the methods proposed by van Ginneken et al. [6] have been outperformed. Our best proposed architecture outperformed Hybrid Voting by almost 8% in Jaccard overlap score. All our architectures performed better than the Original U-Net architecture on all image resolutions. In addition, as one can see from the Table I, results for higher resolutions are much better for smaller objects such as clavicles. Except for InvertedNet architecture which showed a poor performance due to the delayed subsample pooling and small filter sizes in the convolutional layers. On lower resolutions though the InvertedNet demonstrated the best performance on the clavicle segmentation where Original U-Net was surpassed by more

than 7% and the other two networks by 5% and 6% respectively. The performance of the InvertedNet architecture on high resolution can be improved simply by replacing normal convolutional layers by dilated convolutional layers instead. An example of such modification is shown in Fig. 1d and the performance results is added in the Table I. In summary, clavicles are more challenging for Original U-Net, All-Convolutional and All-Dropout on lower resolutions because of the multiple pooling layers in the contractive part of the network and the lack of regularization. In this case the features extracted by the network become less expressive for smaller objects such as clavicles.

Heart segmentation was a challenging task for the InvertedNet architecture. It was even slightly outperformed by the Original U-Net which in its turn was surpassed by the other proposed architectures. Two other proposed architectures All-Convolutional and All-Dropout slightly surpassed the human observer on this task.

The performance of the overall best architecture InvertedNet has been evaluated with several splits of input data into training and testing sets. Table III shows testing results of the InvertedNet trained with the pixelwise cross-entropy loss function. As theoretically expected overall scores get improved when more training data is given to the network. On the other hand increasing difference between numbers of samples in training and testing sets leads to a slight overfitting on the data and therefore increasing of the final generalization error. This is not the case for the negative dice loss function though where increasing the number of training samples gives much better results. Evaluation results for different testing splits for the negative dice loss function are shown in the Table IV.

Fig. 2 shows how the Original U-Net and the proposed models perform on the test set at each epoch during training. The scores of the Original U-Net typically grow faster than the other networks in the beginning but then reach a plateau and oscillate till the end of the training procedure. Other better regularized architectures though start off slower, reach higher or similar scores in the end. InvertedNet starts really slow in the beginning but reaches the best result in the end.

Fig. 3 shows performance results for Original U-Net and the three proposed architectures on the clavicle segmentation task. The X-axis corresponds to intervals for Jaccard scores and the Y-axis corresponds to percentages of test samples falling into the Jaccard intervals from the X-axis. The factor plot has been produced using results for 50%-50% testing split and pixelwise cross-entropy function. InvertedNet has most samples in the last interval and is the only architecture which did not perform worse than 0.7. Jaccard scores for more than a half of the testing samples for InvertedNet reached scores greater than 0.9 which is a proof of the model robustness.

All our proposed architectures reached the best symmetric distance to surface scores among all methods on all organs which is a clear message that convolutional networks are very efficient in extracting features corresponding to object borders. Even in case of quite low contrast difference, for example, on the borders between heart and lung or clavicles and lung.

Fig. 4 shows few examples of the algorithm results for both successful and failed cases. In the white boxes Jaccard

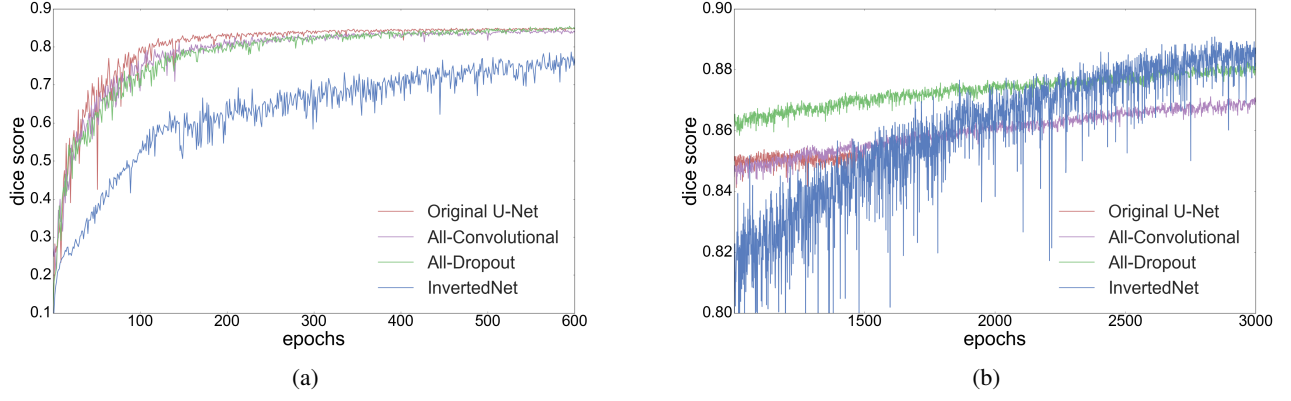


Fig. 2: Performance of models on the test set during training for clavicles a) during early epochs b) during late epochs

Architecture	# of parameters	CPU (s)	GPU (s)	Pros	Size
InvertedNet	3 140 771	7.1	0.06	small # of parameters, best performance	12 MB
All-Convolutional	34 512 388	4.2	0.03	only convolutional layers, no fixed pooling	131 MB
All-Dropout	31 377 988	4.1	0.03	fastest model, nice fit for transfer learning	119 MB

TABLE II: Overview of the proposed architectures

Bodypart	Lungs			Clavicles			Heart		
Evaluation Metric	D	J	S _{sg}	D	J	S _{sg}	D	J	S _{sg}
Training 90%, Testing 10%	0.970	0.943	0.75	0.901	0.820	0.98	0.937	0.881	1.90
Training 70%, Testing 30%	0.970	0.943	0.81	0.899	0.817	1.87	0.934	0.876	1.99
Training 50%, Testing 50%	0.972	0.945	0.76	0.896	0.812	1.52	0.928	0.865	2.07
Training 30%, Testing 70%	0.968	0.939	0.85	0.866	0.764	2.56	0.924	0.859	2.20
Training 10%, Testing 90%	0.948	0.901	1.01	0.805	0.674	2.88	0.875	0.778	2.96

TABLE III: Evaluation comparison of InvertedNet architecture for different test splits for pixelwise cross-entropy loss function

Bodypart	Lungs			Clavicles			Heart		
Evaluation Metric	D	J	S _{sg}	D	J	S _{sg}	D	J	S _{sg}
Training 90%, Testing 10%	0.969	0.941	1.15	0.928	0.866	0.58	0.940	0.887	2.33
Training 70%, Testing 30%	0.971	0.944	0.96	0.931	0.871	0.55	0.932	0.872	2.56
Training 50%, Testing 50%	0.972	0.946	0.73	0.920	0.853	1.10	0.932	0.873	1.98
Training 30%, Testing 70%	0.961	0.925	1.41	0.906	0.828	0.65	0.914	0.842	3.31
Training 10%, Testing 90%	0.940	0.887	2.06	0.837	0.719	1.30	0.875	0.778	4.30

TABLE IV: Evaluation comparison of InvertedNet architecture for different test splits for negative dice loss function

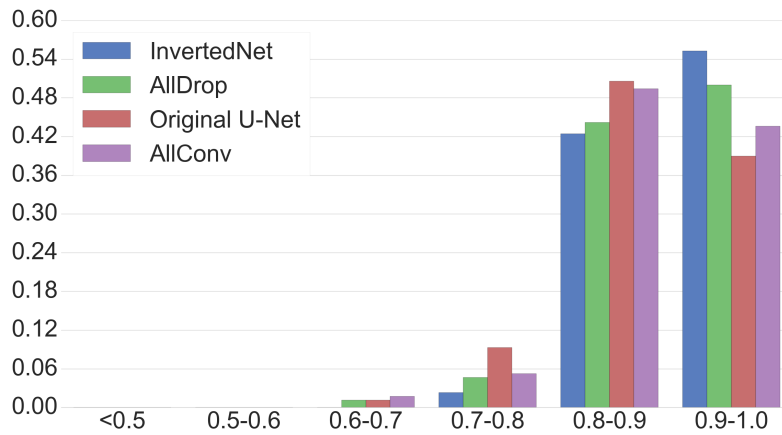


Fig. 3: The percentage contribution (Y-axis) of each model (columns) to each range of Dice score (X-axis) on the test set

scores for lungs, clavicles and heart are shown. To extract the shape contours of the segmentation and ground-truth we used morphological outline extraction algorithm on both segmentation result and reference masks. The contour of the

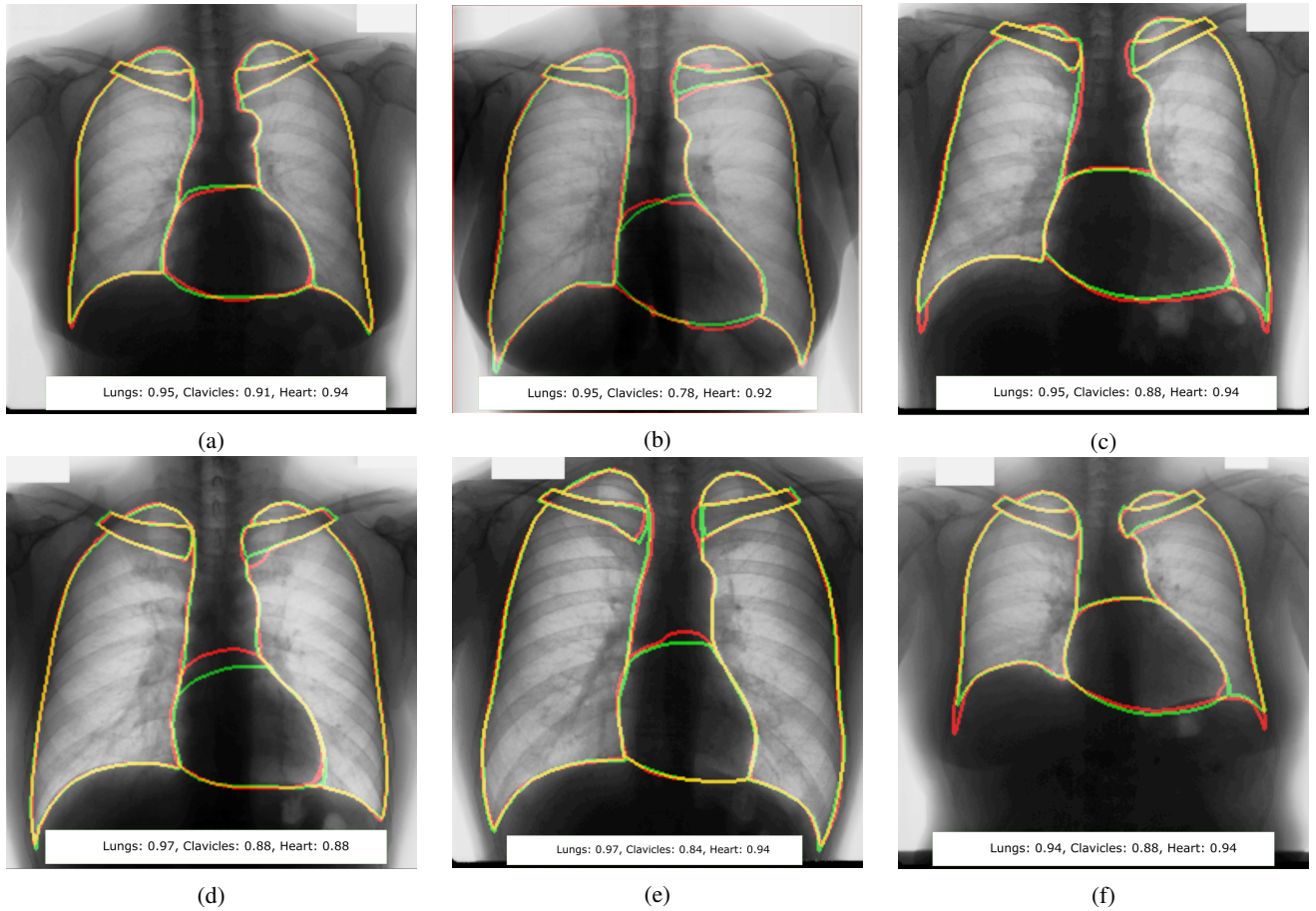


Fig. 4: Segmentation results and corresponding Jaccard scores on some images. The contour of the ground-truth is shown in green, segmentation result of the algorithm in red and overlap of two contours in yellow.

ground-truth is shown in green, segmentation result of the algorithm in red and overlap of two contours in yellow colors respectively.

B. Timing Performance

Table II shows overview of the proposed architectures with execution times measured on the PC with Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30 GHz CPU and GeForce GTX TitanX GPU with 12 GB of memory.

To the best of our knowledge our method is the fastest segmentation approach for CXR images up to date. With modern hardware it can process thousands of images per day which is especially beneficial in big clinical environments when hundreds or sometimes thousands of people are being checked every day.

V. CONCLUSIONS

In this paper we propose an end-to-end approach for multi-class segmentation of anatomical organs in CXR scans. We introduce and evaluate in total three fully-convolutional architectures which reach high test scores on JSRT public dataset showing similar or outperforming the Original U-Net architecture and other state-of-the-art methods on all organs. Our best architecture outperforms the human observer results

on lungs and heart. Clavicle segmentation is still a challenging task for our architectures though they got much closer to the the human observer scores than any other state-of-the-art method did previously. Overall results show that simply adding more regularization and extracting larger number of high level abstract features can be beneficial for improving segmentation of smaller objects such as clavicles. Introducing weighting into the loss-function is crucial when dealing with the highly imbalanced data which is the case in CXR images. Our best architecture has nearly ten times less parameters than the Original U-Net and despite that it outperforms it by a large margin. Trained with the negative Dice loss function it was able to reach mean Jaccard overlap scores of 94.1% for lungs, 86.6% for heart and 88.7% for clavicles.

REFERENCES

- [1] S. Sandström. The who manual of diagnostic imaging, radiographic technique and projections. [Online]. Available: http://www.who.int/diagnostic_imaging/publications/dim_radiotech/en/
- [2] Diagnostic imaging dataset statistical release. [Online]. Available: <https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2015/08/Provisional-Monthly-Diagnostic-Imaging-Dataset-Statistics-2016-07-21.pdf>
- [3] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.

- [4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.
- [5] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, 2000.
- [6] B. van Ginneken, M. B. Stegmann, and M. Loog, "Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database," *Medical Image Analysis*, vol. 10, pp. 19–40, 2006.
- [7] T. Tanimoto, *An Elementary Mathematical Theory of Classification and Prediction*. International Business Machines Corporation, 1958. [Online]. Available: <https://books.google.at/books?id=yp34HAAACAAJ>
- [8] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 577–590, Feb 2014.
- [9] S. Candemir, K. Palaniappan, and Y. Sinan, "Multi-class regularization parameter learning for graph cut image segmentation," in *International Symposium on Biomedical Imaging (ISBI 2013)*, 2013.
- [10] Y. Shao, Y. Gao, Y. Guo, Y. Shi, X. Yang, and D. Shen, "Hierarchical lung field segmentation with joint shape and appearance sparse learning," *IEEE Transactions on Medical Imaging*, vol. 33, no. 9, pp. 1761–1780, Sept 2014.
- [11] B. Ibragimov, B. Likar, F. Pernu, and T. Vrtovec, "Accurate landmark-based segmentation by incorporating landmark misdetections," in *International Symposium on Biomedical Imaging (ISBI 2016)*, April 2016, pp. 1072–1075.
- [12] T. e. a. Xu, "An edge-region force guided active shape approach for automatic lung field detection in chest radiographs," *Computerized Medical Imaging and Graphics*, vol. 36, pp. 452–463, 2012.
- [13] L. e. a. Hogeweg, "Clavicle segmentation in chest radiographs," *Medical Image Analysis*, vol. 16, pp. 1490–1502, 2012.
- [14] H. Boussaid, I. Kokkinos, and N. Paragios, "Discriminative learning of deformable contour models," in *International Symposium on Biomedical Imaging (ISBI 2014)*, April 2014, pp. 624–628.
- [15] S. Candemir, S. Jaeger, W. Lin, Z. Xue, S. Antani, and G. Thoma, "Automatic heart localization and radiographic index computation in chest x-rays," 2016.
- [16] M. Lai, "Deep learning for medical image segmentation," *arXiv e-prints, arXiv:1505.02000*, vol. abs/1505.02000, 2015. [Online]. Available: <http://arxiv.org/abs/1505.02000>
- [17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Conference on Neural Information Processing Systems (NIPS 2012)*, 2012.
- [18] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, May 2016.
- [19] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv e-prints, arXiv:1412.7062v4 [cs.CV]*, vol. abs/1412.7062, 2014. [Online]. Available: <http://arxiv.org/abs/1412.7062>
- [20] O. Tsujii, M. T. Freedman, and S. K. Mun, "Automated segmentation of anatomic regions in chest radiographs using an adaptive-sized hybrid neural network," *Medical physics*, vol. 25, pp. 998–1007, 1998.
- [21] C. Cernazanu-Glavan and S. Holban, "Segmentation of bone structure in x-ray images using convolutional neural network," *Advances in Electrical and Computer Engineering*, vol. 13, no. 1, pp. 87–94, Feb 2013.
- [22] T. A. Ngo and G. Carneiro, "Lung segmentation in chest radiographs using distance regularized level set and deep-structured learning and inference," in *International Conference on Image Processing (ICIP 2015)*, Sept 2015, pp. 2140–2143.
- [23] G. Carneiro, J. C. Nascimento, and A. Freitas, "The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 968–982, March 2012.
- [24] M. Havaci, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical Image Analysis*, 2016.
- [25] P. Petersen, M. Nielsen, P. Diao, N. Karssemeijer, and M. Lillholm, *Breast tissue segmentation and mammographic risk scoring using deep learning*. Springer Science+Business Media B.V., 2014, pp. 88–94.
- [26] B. Gaonkar, D. Hovda, N. Martin, and L. Macyszyn, "Deep learning in the small sample size setting: cascaded feed forward neural networks for medical image segmentation," pp. 97 852I–97 852I–8, 2016.
- [27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR 2015)*, 2015.
- [28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional Networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*. Springer, 2015, pp. 234–241.
- [29] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [30] X. Bouthillier, K. Konda, P. Vincent, and R. Memisevic, "Dropout as data augmentation," *arXiv e-prints, arXiv:1412.7062v4 [cs.CV]*, vol. 1050, p. 8, 2016.
- [31] P. Gravel, G. Beaudoin, and J. A. De Guise, "A method for modeling noise in medical images," *IEEE Transactions on Medical Imaging*, vol. 23, no. 10, pp. 1221–1232, 2004.
- [32] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *International Conference on Learning Representations (ICLR 2015)*, 2014.
- [33] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, pp. 947–951, 2000.
- [34] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR 2015)*, 2015.